Journal of Physics: Conference Series 2014 vol.574 N1

# Cluster analysis of word frequency dynamics

Maslennikova Y., Bochkarev V., Belashova I.
*Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia*

## Abstract

© Published under licence by IOP Publishing Ltd. This paper describes the analysis and modelling of word usage frequency time series. During one of previous studies, an assumption was put forward that all word usage frequencies have uniform dynamics approaching the shape of a Gaussian function. This assumption can be checked using the frequency dictionaries of the Google Books Ngram database. This database includes 5.2 million books published between 1500 and 2008. The corpus contains over 500 billion words in American English, British English, French, German, Spanish, Russian, Hebrew, and Chinese. We clustered time series of word usage frequencies using a Kohonen neural network. The similarity between input vectors was estimated using several algorithms. As a result of the neural network training procedure, more than ten different forms of time series were found. They describe the dynamics of word usage frequencies from birth to death of individual words. Different groups of word forms were found to have different dynamics of word usage frequency variations.

http://dx.doi.org/10.1088/1742-6596/574/1/012120